04109520 **Image available**

CHARACTER RECOGNIZER

PUB. NO.: 05-101220 [JP 5101220 A]
PUBLISHED: April 23, 1993 (19930423)
INVENTOR(s): FURUTO TAKESHI
APPLICANT(s): SUMITOMO ELECTRIC IND LTD [000213] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 03-260659 [JP 91260659]
FILED: October 08, 1991 (19911008)
INTL CLASS: [5] G06K-009/34; G06F-015/20
JAPIO CLASS: 45.3 (INFORMATION PROCESSING -- Input Output Units); 45.4 (INFORMATION PROCESSING -- Computer Applications)
JAPIO KEYWORD:R107 (INFORMATION PROCESSING -- OCR & OMR Optical Readers)
JOURNAL: Section: P, Section No. 1596, Vol. 17, No. 454, Pg. 97, August 19, 1993 (19930819)

ABSTRACT

PURPOSE: To recognize the characters at a high speed by segmenting these characters with high accuracy and at a high speed.
CONSTITUTION: The images are segmented for each line (n2) and then the images of each character element, i.e., the groups of picture elements are segmented (n3). The character elements having the small lateral widths are extracted as the half size candidate characters based on the character size estimated from the size of the character element (n6). Then the half size candidate character strings that satisfy the fixed conditions are extracted as the English word candidates (n7). The picture element features, etc., are extracted in regard of each half size candidate character forming an English word candidate (n8). Based on these picture element features, etc., the English word candidates including the half size candidate characters which are not identical with the half size English letters are deleted out of all. English word candidates and the English word candidates are corrected (n9). Then the segmenting position is corrected in accordance with the half end full size characters respectively (n10). Then the characters are recognized (n11).

DIALOG(R)File 345:Inpadoc/Fam.& Legal Stat

11157512
Basic Patent (No,Kind,Date): JP 5101220 A2 930423 <No. of Patents: 001>
Patent Family:

Patent No	Kind	Date	Applic No	Kind	Date	
JP 5101220	A2	930423	JP 91260659	A	911008	(BASIC)

Priority Data (No,Kind,Date):
JP 91260659 A 911008

PATENT FAMILY:
JAPAN (JP)
Patent (No,Kind,Date): JP 5101220 A2 930423
CHARACTER RECOGNIZER (English)
Patent Assignee: SUMITOMO ELECTRIC INDUSTRIES
Author (Inventor): FURUTO TAKESHI
Priority (No,Kind,Date): JP 91260659 A 911008
Applic (No,Kind,Date): JP 91260659 A 911008
IPC: * G06K-009/34; G06F-015/20
JAPIO Reference No: ; 170454P000097
Language of Document: Japanese

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平5－101220

(43)公開日　平成５年(1993)４月23日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｋ　9/34		9073－5L		
Ｇ０６Ｆ　15/20	５９２　Ｄ	7343－5L		

審査請求　未請求　請求項の数３(全 16 頁)

(21)出願番号　特願平3－260659

(22)出願日　平成３年(1991)10月８日

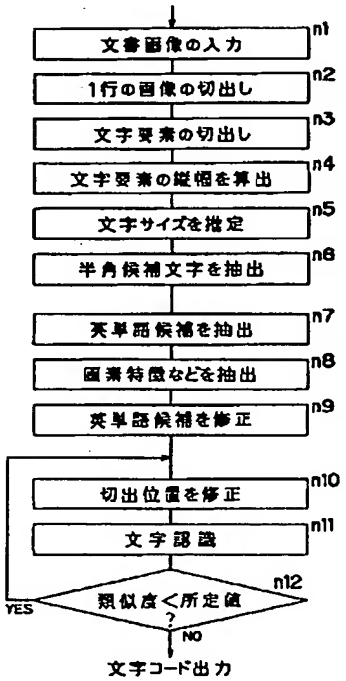
(71)出願人　000002130
住友電気工業株式会社
大阪府大阪市中央区北浜四丁目５番33号

(72)発明者　古戸　健
大阪市此花区島屋一丁目１番３号　住友電気工業株式会社大阪製作所内

(74)代理人　弁理士　亀井　弘勝　(外２名)

(54)【発明の名称】　文字認識装置

(57)【要約】
【目的】正確かつ高速な文字切出しを実現して文字認識処理を高速化する。
【構成】１行毎の画像の切出しが行われ（ステップｎ２）、さらに画素の塊である文字要素毎の画像が切り出される（ステップｎ３）。次に、文字要素の大きさから推定された文字サイズに基づき、横幅が小さい文字要素が半角候補文字として抽出される（ステップｎ６）。さらに、一定の条件を満たす半角候補文字列が英単語候補として抽出される（ステップｎ７）。次いで、英単語候補を構成する各半角候補文字に関して、画素特徴などが抽出される（ステップｎ８）。この画素特徴などに基づき、半角英文字でない半角候補文字を含む英単語候補が英単語候補から除外されて、英単語候補の修正が行われる（ステップｎ９）。そして、切出位置が半角文字と全角文字とのそれぞれに応じて修正され（ステップｎ１０）、文字認識処理が行われる（ステップｎ１１）。



1

【特許請求の範囲】

【請求項１】日本語文字および英文字が混在している文書画像の各文字を認識し、文字コードに変換して出力する文字認識装置であって、

入力画像の行方向の周辺分布をとり、１行ずつの画像を切り出す行切出手段と、

この行切出手段により切り出された１行毎の画像について、行方向に垂直な方向に関する周辺分布をとり、文字を構成する画素の塊である文字要素毎の画像を切り出す仮切出手段と、

切り出された文字要素の大きさに基づいて日本語文字の大きさである文字サイズを推定する文字サイズ推定手段と、

この文字サイズ推定手段により推定された文字サイズに基づいて、横幅が文字サイズの一定割合よりも小さな文字要素を半角候補文字として抽出する半角候補文字抽出手段と、

所定数以上の半角候補文字が連続し、かつ、この半角候補文字列の前または後ろに所定長以上の余白部分が存在するときに、当該半角候補文字列を英単語候補として抽出する英単語候補抽出手段と、

英単語候補を構成する各半角候補文字の行内での位置特徴、および当該半角候補文字の構成画素の分布状態に対応した画素特徴を抽出する画素特徴抽出手段と、

半角英文字に関して、上記位置特徴および画素特徴についての標準条件を記憶した画素特徴記憶手段と、

上記画素特徴抽出手段により抽出された上記位置特徴および画素特徴と、上記特徴記憶手段に記憶された標準条件とを照合して、位置特徴および画素特徴が上記標準条件に合致しないときに、当該半角候補文字を含む英単語候補を英単語候補から除外する英単語候補修正手段と、

英単語候補を構成する半角候補文字は半角英文字であるものとして文字画像の切出しを行い、残余の半角候補文字は日本語文字の一部を成すものとして近傍の半角候補文字と再結合させて文字画像の切出しを行う文字切出手段と、

この文字切出手段により切り出された画像に基づいて文字認識を行い、対応する文字コードを出力する認識手段とを含むことを特徴とする文字認識装置。

【請求項２】上記画素特徴には、半角候補文字を行方向の中心位置で行方向に垂直な方向に走査したときに、白画素→黒画素→白画素のように変化する変化点の数、または黒画素→白画素→黒画素のように変化する変化点の数が含まれていることを特徴とする請求項１記載の文字認識装置。

【請求項３】上記認識手段は、

認識可能な全ての文字の特徴量を記憶した認識用辞書と、

上記切り出された文字画像から特徴量を抽出する特徴量抽出手段と、

2

この特徴量抽出手段により抽出された特徴量と上記認識用辞書に記憶されている各文字の特徴量との類似度を算出する類似度算出手段と、

この類似度算出手段により算出された類似度が最大である文字の文字コードおよびその類似度を出力する手段とを含むものであり、

上記出力された類似度が所定値未満であるときには、当該文字要素について上記文字切出手段による文字画像の切出しとは異なる態様で文字画像の切出しを再度行う再切出手段をさらに含むことを特徴とする請求項１または２記載の文字認識装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、光学的文字読取装置（ＯＣＲ）などのように、文書画像から１文字ずつの画像を切り出し、この切り出した画像に基づいて文字や記号の認識処理を行う文字認識装置に関するものである。

【０００２】

【従来の技術】光学的文字読取装置（ＯＣＲ）などの文字認識装置では、文書画像に対応した画像データから、１文字分ずつの画像を順に切り出し、切り出された画像の特徴量と、認識用辞書に蓄積されている認識可能な各文字の特徴量との比較演算を行うことで、文字の認識が行われ、認識された文字に対応した文字コードが出力される。

【０００３】このような文字認識処理が正確に行われるためには、文書画像からの１文字ずつの文字切出しが良好に行われることが不可欠である。この文字切出処理に当たっては、まず文字サイズを検出することが必要となる。これは、たとえば「い」や「旧」などのように分離した複数の要素から構成されている分離文字では、黒画素の塊である文字要素毎に画像を切り出すと、正しい文字切出処理が行えないからである。すなわち、予め文字サイズを検出しておき、複数の文字要素からなる分離文字に対しては、文字サイズに基づいて複数の文字要素を再構成することにより文字切出処理を正確に行える。

【０００４】しかしながら、日本語の文書中に半角英数字が存在する場合などには、一定の文字サイズを基に文字の切出しを行うと、隣接する半角英数字の対が１文字として再構成されて切り出されるおそれがある。このため、日本語と半角英数字とが混在している文書では、単純にサイズの小さな文字要素同士を結合させてしまうと、正確な文字切出処理を行うことができない。

【０００５】この不具合を解消するために、文字要素毎の認識結果をフィードバックして、正しい切出位置を定める方法が提案されている（たとえば信学論 '84/10 Vol.J67-D No.10 参照）。ところが、この方法では、全角分離文字（横書きの文書では、行方向に関して白画素で分離された複数の文字要素からなる文字。たとえば「門」や「卵」などが該当する。）に関しても、各文字

要素毎に認識が行われるため、全体の文字認識回数が文書の構成文字数に比較して増大してしまい、処理速度が劣化するという問題が生じていた。

【0006】すなわち、たとえば、図16(a)および(b)に示すような「門」や「卵」のような全角分離文字に関しては、参照符号11，12で切出位置を示すように、それぞれ2つずつの文字要素A1，A2；B1，B2毎に切り出されるので、先ず各文字要素A1，A2；B1，B2に対して認識処理が行われる。そして、この認識処理が不可能であることに基づいて、文字要素A1，A2またはB1，B2を結合させて参照符号13，14で示す位置で切出しをやり直し、このようにして切り出された文字要素の対に関してさらに文字認識処理を行うことになる。したがって、本来2回の認識処理により認識すべき2つの文字に対して6回の認識処理を要することになる。

【0007】そこで、特開昭62－130479号公報に開示されている他の先行技術では、推定文字ピッチを基に判定された基準文字塊（全角文字に対応する。）に挟まれた小さな文字要素の数に応じて処理を異ならせることにより、処理速度の向上が図られている。すなわち、小さな文字要素が所定数（たとえば4個）以上連続するときには、各文字要素が1文字（半角文字）を構成するものとして認識処理が行われる。そして、小さな文字要素の連続数が所定数未満であるときには、分離文字（全角分離文字）の可能性があるものとして、各文字要素毎の認識処理を行うとともに、2個以上の文字要素の結合に関しても認識処理を行わせるようにしている。このような処理により、半角英数字が連続する場合には、このような文字列については、文字要素の結合についての認識処理が省かれるから、認識処理の回数を低減して、処理速度を向上することができる。

【0008】この先行技術の新たな問題は、図17に示すように、分離文字が連続する場合に、分離文字の各文字要素をそれぞれ半角文字であるものとして認識処理が行われるために、認識処理のやり直しが必要となり、結果的に却って処理速度が落ちてしまう場合があることである。すなわち、図17の場合のように、それぞれ文字要素C11，C12；C21，C22；C31，C32からなる分離文字C1，C2，C3が連続する場合には、小さな文字要素C11，C12，C21，C22，C31，C32が6個連続しているから、これらの文字要素に対しては、文字要素相互間の結合は行われず、先ず、半角文字であるものとして認識処理が行われることになる。そして、この半角文字としての認識が不可能であると判った時点で、初めて、隣接する文字要素間の再結合が行われる。なお、図17において、15は行方向の切出位置を示す。

【0009】上記の不具合を解決するためには、半角文字と分離文字を構成する小さな文字要素とを、認識処理を行うことなく区別する必要がある。これを実現した先行技術は、たとえば特開平2－239386号公報に開示されている。この先行技術では、認識処理を行う前に、文字要素の行内における相対的な位置関係や文字要素の大きさを基にして、当該文字要素が半角文字を構成しているのか、分離文字の一部であるのかを判断するようにしている。

【0010】

【発明が解決しようとする課題】ところが、この先行技術は、文字要素の位置関係やレイアウトに関する規則を用いて文字要素の判別を行っているため、文字の切出しの正確さが文書フォーマットの影響を受け易いという新たな問題を生じさせる。そこで、本発明の目的は、上述の技術的課題を解決し、文書フォーマットの影響を受けることなく、日本語文書中に半角英文字が混在している文書画像からの文字の切出しを正確に、かつ、高速に行うことができ、したがって、文字認識処理の高速化に寄与することができる文字認識装置を提供することである。

【0011】

【課題を解決するための手段および作用】上記の目的を達成するための請求項1記載の文字認識装置は、日本語文字および英文字が混在している文書画像の各文字を認識し、文字コードに変換して出力する文字認識装置であって、入力画像の行方向の周辺分布をとり、1行ずつの画像を切り出す行切出手段と、この行切出手段により切り出された1行毎の画像について、行方向に垂直な方向に関する周辺分布をとり、文字を構成する画素の塊である文字要素毎の画像を切り出す仮切出手段と、切り出された文字要素の大きさに基づいて日本語文字の大きさである文字サイズを推定する文字サイズ推定手段と、この文字サイズ推定手段により推定された文字サイズに基づいて、横幅が文字サイズの一定割合よりも小さな文字要素を半角候補文字として抽出する半角候補文字抽出手段と、所定数以上の半角候補文字が連続し、かつ、この半角候補文字列の前または後ろに所定長以上の余白部分が存在するときに、当該半角候補文字列を英単語候補として抽出する英単語候補抽出手段と、英単語候補を構成する各半角候補文字の行内での位置特徴、および当該半角候補文字の構成画素の分布状態に対応した画素特徴を抽出する画素特徴抽出手段と、半角英文字に関して、上記位置特徴および画素特徴についての標準条件を記憶した画素特徴記憶手段と、上記画素特徴抽出手段により抽出された上記位置特徴および画素特徴と、上記特徴記憶手段に記憶された標準条件とを照合して、位置特徴および画素特徴が上記標準条件に合致しないときに、当該半角候補文字を含む英単語候補を英単語候補から除外する英単語候補修正手段と、英単語候補を構成する半角候補文字は半角英文字であるものとして文字画像の切出しを行い、残余の半角候補文字は日本語文字の一部を成すもの

として近傍の半角候補文字と再結合させて文字画像の切出しを行う文字切出手段と、この文字切出手段により切り出された画像に基づいて文字認識を行い、対応する文字コードを出力する認識手段とを含むものである。

【００１２】上記の構成によれば、行切出手段により切り出された１行の文書画像から、文字を構成する画素の塊である文字要素が仮切出手段によって切り出され、この切り出された文字要素の大きさに基づいて当該文書画像を構成する日本語文字の大きさである文字サイズが、文字サイズ推定手段によって推定される。そして、推定された文字サイズに基づいて、半角候補文字抽出手段では、横幅が文字サイズの一定割合よりも小さな文字要素である半角候補文字が抽出される。さらに、この半角候補文字が所定数以上連続し、かつ、この所定数以上の一連の半角候補文字からなる半角候補文字列の前または後ろに所定長以上の余白部分が存在するときに、半角候補文字列が英単語候補として抽出される。

【００１３】すわなち、日本語文書中に存在する半角英文字は、一般に、単語を形成しており、しかも、前後に余白が設けられている場合が多い。英単語候補文字抽出手段での英単語候補文字の抽出は、日本語文書中の半角英文字の上記のような特徴を利用したものである。英単語候補が抽出されると、この英単語候補を構成する半角候補文字に関して、行内での位置特徴と、当該半角候補文字の構成画素の分布特徴に対応した画素特徴が抽出される。一方、位置特徴および画素特徴についての標準条件は、画素特徴記憶手段に記憶されており、英単語候補修正手段は、画素特徴抽出手段により抽出された位置特徴および画素特徴を上記の標準条件と照合し、この標準条件に合致しない半角候補文字を含む英単語候補を英単語候補から除外する。

【００１４】すなわち、英単語候補抽出手段での英単語候補抽出処理は、文書画像のフォーマットの影響を受けやすいため、必ずしも確実に英単語候補を抽出できるとは限らない。そこで、個々の半角候補文字の位置特徴および画素特徴に基づいて各半角候補文字が英文字である可能性を調べ、いずれか１つでも英文字ないものと判断される半角候補文字を含む英単語候補を英単語候補から除外することとしている。このようにして、英単語候補を正確に抽出することが可能となる。

【００１５】このように、英単語候補を構成する半角候補文字の画素特徴を参照して英単語候補の修正を行うこととしている結果、種々のフォーマットの文書に良好に対応して、英単語候補を確実に抽出できる。これににより、文書フォーマットに依らずに良好に切出処理を行わせて、文字認識処理を良好に行わせることができる。文字切出手段では、英単語候補を構成する半角候補文字は半角英文字であるものとして文字画像の切出しが行われ、それ以外の半角候補文字は、日本語文字の一部をなすものであるとして近傍の半角候補文字と再結合されて

切り出される。これにより、半角英文字は確実に半角文字として切り出すことができ、また左右に分離している全角分離文字については、再結合を行わせて切り出させることができる。

【００１６】この結果、全角分離文字が半角文字として切り出されて認識処理が行われたり、半角英文字が全角文字として切り出されて認識処理が行われたりすることを有効に防ぐことができるから、文字認識処理の回数を低減して、処理の高速化に寄与することができる。なお、上記画素特徴には、半角候補文字を行方向の中心位置で行方向に垂直な方向に走査したときに、白画素→黒画素→白画素のように変化する変化点の数、または黒画素→白画素→黒画素のように変化する変化点の数が含まれていてもよい。

【００１７】請求項３記載の文字認識装置は、上記認識手段は、認識可能な全ての文字の特徴量を記憶した認識用辞書と、上記切り出された文字画像から特徴量を抽出する特徴量抽出手段と、この特徴量抽出手段により抽出された特徴量と上記認識用辞書に記憶されている各文字の特徴量との類似度を算出する類似度算出手段と、この類似度算出手段により算出された類似度が最大である文字の文字コードおよびその類似度を出力する手段とを含むものであり、上記出力された類似度が所定値未満であるときには、当該文字要素について上記文字切出手段による文字画像の切出しとは異なる態様で文字画像の切出しを再度行う再切出手段をさらに含むことを特徴とする。

【００１８】この構成では、文字コードとともに出力される類似度が所定値未満であるときには、そのような文字要素については再切出手段によって、文字画像の切出しが再度行われる。すなわち、半角文字として切り出して認識したが類似度が低い場合には、近傍の半角候補文字と再結合させて再切出しが行われ、また全角文字として切り出して認識したが類似度が低い場合には、半角文字として切出しが再度行われる。これにより、認識誤りを低減して、確実な文字認識を行うことができる。

【００１９】

【実施例】以下実施例を示す添付図面によって詳細に説明する。図２は本発明の文字認識装置の一実施例である光学的文字読取装置の基本的な構成を示すブロック図である。原稿１の表面に形成された文書画像は、イメージスキャナ２により光学的に読み取られる。原稿１には、日本語の文書が形成されており、この文書中には、半角英文字で構成された英単語が含まれているものとする。イメージスキャナ２の出力信号は、二値化部３で二値化された後に、二値画像として画像メモリ４に記憶される。

【００２０】画像メモリ４に記憶された文書画像は、切出部５で１文字ずつの文字画像毎に切り出される。そして、切り出された画像に基づき、認識部６において文字

7

認識処理が行われ、当該文字に対応する文字コードが出力される。図1は、切出部5の詳しい構成を示すブロック図である。画像メモリ4に記憶れた文書画像は、行切出部51において1行ずつの画像ごとに切り出される。この切り出された1行毎の画像は、仮切出部52に与えられ、黒画素の塊である文字要素毎の画像に切り出される。

【0021】切り出された文字要素は、文字サイズ推定部53に与えられ、公知の文字サイズ推定手法に基づいて、日本語文字のサイズが推定される。この推定された文字サイズは、半角候補文字抽出部54に与えられて、文字サイズの一定割合以下の横幅の文字要素が、半角候補文字として抽出される。抽出された半角候補文字は、英単語候補抽出部55に与えられる。この英単語候補抽出部55は、後述する手法によって、一定の条件を満たす半角候補文字列を英単語候補として抽出する。

【0022】英単語候補が抽出されると、次に、英単語候補を構成する各半角候補文字に関して、行内での位置特徴や、各半角候補文字の構成画素の分布の特徴である画素特徴が、画素特徴抽出部56で抽出される。抽出された画素特徴などは、照合部57において、画素特徴記憶部58に記憶された画素特徴テーブルの内容と比較照合される。この画素特徴テーブルには、半角英文字について、上記の位置特徴および画素特徴に関する標準条件が記憶されている。

【0023】照合部57での照合結果は、英単語候補修正部59に与えられる。この英単語候補修正部59は、上記の照合結果に基づいて、英単語候補を構成する半角候補文字の少なくともいずれか1つが英文字でないものと判断されるときに、このような半角候補文字を含む英単語候補を英単語候補から除外する。この英単語候補修正部59における詳しい処理については、後述する。

【0024】このようにして英単語候補の修正が行われて最終的に英単語候補が確定すると、文字切出部60では、英単語候補を構成する半角候補文字は、半角英文字であるものとして文書画像からの画像切出を行う。一方、英単語候補を構成しない半角候補文字は、隣接する半角候補文字との再結合が行われて切り出される。すなわち、このような半角候補文字は、全角分離文字の一部をなす文字要素であるものとして切り出されることになる。

【0025】図3は、認識部6の詳しい構成を示すブロック図である。切出部5からの1文字ごとの切出画像は、特徴抽出部61に入力され、種々の特徴量が抽出されることになる。抽出された特徴量は、類似度算出部62に与えられる。この類似度算出部62には、半角英数字の特徴量を記憶した認識用辞書である英数字辞書63と、半角英数字以外の認識可能な全ての文字の特徴量を記憶した認識用辞書64とが接続されている。類似度算出部62では、特徴抽出部62から与えられる特徴量

8

と、辞書63，64に記憶された特徴量との類似度が計算される。そして、最大の類似度を有する文字が見出され、その文字の文字コードとその類似度とが出力部65から出力されることになる。

【0026】なお、上記最大の類似度は切出部5にも与えられ、この類似度が所定値未満のときには、切出ミスであるものと判断されて、当該文字要素について再度の切出処理が行われる。図4は文字認識処理を説明するためのフローチャートである。ステップn1では、画像メモリ4に文書画像が入力される。ステップn2では、行切出部51において、行方向に関して黒画素の周辺分布が計算され、行切出位置および行幅が求められて、1行の画像の切出しが行われる。すなわち、図5に示すように、行方向に関して黒画素数が累積され、参照符号a1で示すヒストグラムが作成される。このヒストグラムに基づいて行幅Vが算出される。さらに、行切出位置L1，L2，L3，‥‥も算出されることになる。この行切出位置L1，L2，L3，‥‥および行幅Vに基づいて、1行の画像の切出しが行われる。

【0027】次に、ステップn3では、仮切出部52において、行切出部51により切り出された各行の画像毎に、行方向に垂直な方向に関して黒画素の周辺分布が求められる。すなわち、図6に示すように、切り出された各行の画像において行方向に垂直な方向に関して黒画素数が累積され、参照符号a2で示すヒストグラムが作成される。そして、このヒストグラムに基づいて、黒画素の塊である各文字要素の切出位置hk（ただし、k＝1，2，3，‥‥である。）および横幅Hkが計算される。このようにして、文字要素の切出しが達成される。

【0028】ステップn4では、図7に示すように各文字要素毎に、行幅Vの範囲内で、文字の縦方向の黒画素の周辺分布b1，b2，b3，‥‥が求められ、各文字要素の縦幅V1，V2，V3，‥‥が算出される。そして、ステップn5では、算出された縦幅V1，V2，V3，‥‥などに基づいて、各文字要素を行方向または行垂直方向に沿う四辺で囲む最小の矩形が求められ、この矩形が囲み切出矩形とされる。さらに、囲み切出矩形の大きさから、文字サイズ推定部53において、日本語文字の大きさである文字サイズの推定が行われる。この文字サイズの推定には、たとえば「横書き日本語文書における個別文字の抽出（信学論'85/11, Vol.J68-D No.11, 第1899頁」などに開示されている公知技術を適用することができる。

【0029】ステップn6では、半角候補文字抽出部54において、囲み切出矩形の大きさと文字サイズとの関係から、半角候補文字が抽出される。この半角候補文字とは、切出幅の推定文字サイズに対する割合が一定値よりも小さな文字要素のことである。この半角候補文字の抽出処理は、図8に示されている。すなわち、隣接する一対の文字要素の結合の横幅W1，W2，W3，‥‥が

9

算出され、この算出された横幅W１，W２，W３，‥‥がステップｎ５で推定された文字サイズと比較される。図８において、実線で示されている横幅W１，W３～W９，W１１は文字サイズ以下であり、破線で示された横幅W２，W１０，W１２，W１３は文字サイズを超えている。この横幅W１，W２，‥‥に基づき、１つの文字要素に関連する２つの横幅Wｊおよび W（ｊ＋１）（ｊ＝１，２，３，‥‥）の少なくともいずれか一方が文字サイズ以下であれば、当該文字要素は半角候補文字とされる。一方、２つの横幅Wｊおよび W（ｊ＋１）のいずれもが文字サイズを超えているときには、当該文字要素は半角候補文字からは除外される。したがって、図８の場合には、文字要素２１～３２は、半角候補文字とされ、文字要素３３，３４は半角候補文字からは除外され、全角文字としての文字切出処理が行われることになる。

【００３０】図９は、半角候補文字の具体的な抽出方法を説明するためのフローチャートである。この処理は、図４のステップｎ２で抽出した各行の画像毎に行われる。ステップｓ１では、処理中の文字要素を計数するためのパラメータｍに１が代入される。ステップｓ２では、変数Wに、m番目の文字要素の横幅HWmと、（m＋１）番目の文字要素の横幅HW（m＋１）と、m番目の文字および（m＋１）番目の文字の間隔Ｓmとの和が代入される。この変数Wが上記の横幅W１，W２，W３，‥‥に対応する。

【００３１】次に、ステップｓ３では、変数Wが、図４のステップｎ５で抽出した文字サイズ以下であるかどうかが判断される。変数Wが文字サイズ以下であるときに、ステップｓ４でm番目および（m＋１）番目の文字要素が半角候補文字とされる。ステップｓ５では、パラメータmがインクリメントされ、ステップｓ６での処理により、パラメータmが（文字要素の数－１）となるまで同様の処理が行われる。ステップｓ３で変数Wが文字サイズを超えているものと判断されるときには、ステップｓ４を経ずにステップｓ５に移る。このような処理によって、隣接する文字要素と結合し得ない文字要素が半＊

10

＊角候補文字として抽出されることになる。

【００３２】再び図４を参照する。半角候補文字は、切り出された文字要素のサイズからは、半角文字であるのか、左右に分離された形態を有する全角分離文字の一部であるのかが判断できない。このため、図４のステップｎ７では、英単語候補抽出部５５において、一定の条件を満たす半角候補文字列が英単語候補として抽出される。

【００３３】一般に、日本語文書中に混在する英単語を構成する半角文字列は、図１０(a)に示すように略等しい間隔S_1，S_2，‥‥，S_6（$S_1 \fallingdotseq S_2 \fallingdotseq \cdots \fallingdotseq S_6$）で印字（以下「等間隔印字」という。）されるか、または、図１０(b)に示すように略等しいピッチP_1，P_2，‥‥，P_6（$P_1 \fallingdotseq P_2 \fallingdotseq \cdots \fallingdotseq P_6$）で印字（以下「等ピッチ印字」という。）される。等ピッチ印字では、間隔S_1，S_2，‥‥，S_6のばらつきが大きいのが通常である。したがって、半角候補文字がほぼ一定間隔またはほぼ一定のピッチで一定数以上連なっていれば、このような半角候補文字列は英単語を構成している可能性が高い。

【００３４】さらに、日本後文書中に混在する英単語は、その前後にある程度の余白が設けられるという特徴がある。そこで、本実施例では、図１１に示すように、半角候補文字がほぼ一定間隔または一定ピッチで一定数（たとえば４個）以上連なることを第１の条件とし、そのようなひとつながりの半角候補文字列のうちで一番後ろに位置する半角候補文字とその次に位置する文字要素との間隔が大きく開いていることを第２の条件として、この第１および第２の条件を満たす半角候補文字列が英単語候補として抽出される。

【００３５】たとえば、図１１に示す例では、半角候補文字Ｃ２１～Ｃ２８は、ピッチP_{21}，P_{22}，P_{23}，P_{24}，P_{25}，P_{26}，P_{27}で並んでおり、また、各半角候補文字Ｃ２１～Ｃ２８間の各間隔は、S_{21}，S_{22}，S_{23}，S_{24}，S_{25}，S_{26}，S_{27}となっている。この場合、

$$P_{21} \fallingdotseq P_{22} \fallingdotseq P_{23} \fallingdotseq P_{24} \fallingdotseq P_{25} \fallingdotseq P_{26} \fallingdotseq P_{27} \ll P_{28} \quad \cdots\cdots (1)$$

または

$$S_{21} \fallingdotseq S_{22} \fallingdotseq S_{23} \fallingdotseq S_{24} \fallingdotseq S_{25} \fallingdotseq S_{26} \fallingdotseq S_{27} \ll S_{28} \quad \cdots\cdots (2)$$

が成り立つから、半角候補文字Ｃ２１～Ｃ２７は、ほぼ一定ピッチまたはほぼ一定ピッチで一定数以上連なっていると言える。すなわち、半角候補文字Ｃ２１～Ｃ２８からなる文字列は、上記の第１の条件を満たす。

【００３６】このようなほぼ一定間隔またはほぼ一定ピッチで並んでいる半角候補文字列を構成する半角候補文字Ｃ２１～Ｃ２７のうち、一番後ろの半角候補文字Ｃ２７の次の間隔S_{28}が大きいから、半角候補文字Ｃ２１～Ｃ２７からなる半角候補文字列は、結局上記第２の条件をも満たす。この結果、この間隔S_{28}より前の半角候補文字列が英単語候補とされることになる。

【００３７】その一方で、キャラクタ「が」を構成する半角候補文字列Ｃ３１，Ｃ３２は、その直後の間隔S_{32}が小さすぎ、また半角候補文字も２つしか連なっていないので、英単語候補とは判定されない。また、キャラクタ「は」を構成する２つの半角候補文字Ｃ１１，Ｃ１２は、後ろの間隔S_{12}は充分大きいが、半角候補文字が２つしか連なっていないから、英単語候補とは判定されない。

【００３８】このようにして、文字列「Ｅｉｎｓｔｅｉ

11

n」を構成する半角候補文字列C21～C27が英単語候補として抽出されることになる。このようにして、英単語候補が抽出されると、次に、図4のステップn8では、画素特徴抽出部56において、図12に示されるように、英単語候補を構成する各半角候補文字の画素特徴などが抽出される。この画素特徴とは、たとえば、各半角候補文字を行方向に関する中心位置で縦方向に走査したときに、白画素→黒画素→白画素のような変化が生じる変化点の数numなどである。各半角候補文字の走査結果は、図12において参照符号d1～d4で示されており、太線部分は走査線30上の画素が黒画素であることを示し、細線部分は走査線30上の画素が白画素であることを示す。

【0039】たとえば、図12(a)において、走査線30に沿って半角候補文字「E」を走査すると、上記のような変化点は参照符号c1，c2，c3で示す3個となる（すなわちnum＝3である。）。同様に、図12(b),(c),(d)から、半角候補文字「i」ではnum＝2、半角候補文字「n」ではnum＝1、半角候補文字「s」ではnum＝3となる。

【0040】なお、図13は、キャラクタ「は」の分離された各文字要素に関して上記と同様な走査を行った結果を参考のために示したものである。図14は画素特徴抽出部56で抽出されるその他の画素特徴を説明するための図である。半角候補文字の上記の走査の際には、走査線30上において、連続した白画素からなる白画素群と連続した黒画素からなる黒画素群との各構成画素数が計数される．すなわち、走査線30に従った走査において、最初に現れる白画素群の画素数W（1）、その次に現れる最初の黒画素群の画素数B（1）、その次に現れる2番目の白画素群の画素数W（2）、その次に現れる2番目の黒画素群の画素数B（2）、‥‥が計数される。

【0041】さらに画素特徴抽出部56では、行内での半角候補文字の相対位置などを表す位置特徴として、行幅V内における半角候補文字の囲み切出矩形35の上部余白SUおよび下部余白SLも求められ、行内での文字の相対位置が調べられる。この余白SUおよびSLは、いずれも、画素数で表されるデータである。これらのデータの他に、さらに、上記図7に示された処理により求められたヒストグラムから、半角候補文字の縦方向の分

12

離数SPが求められ、この分離数SPも位置特徴とされる。すなわち、キャラクタ「d」では、SP＝1であり、キャラクタ「i」では、SP＝2である。

【0042】上記のような画素特徴の抽出結果の一例は、上記の図12に示されている。また、参考のために、図13には、キャラクタ「は」の分離された各文字要素に関して画素特徴を抽出した例が示されている。図1の画素記憶部58には、半角英文字の位置特徴および画素特徴に関する標準条件を格納したテーブルが記憶されている。このテーブルを参照することにより、図4のステップn9では、照合部57および英単語候補修正部50での処理によって、英単語候補を構成する半角候補文字が半角英文字であるかどうかが再度調べられ、英単語候補の修正が行われる。すなわち、上記テーブルと抽出された位置特徴および画素特徴とが照合され、英単語候補とされた文字要素列を構成する半角候補文字のうち、いずれか1つでも英文字に該当しないものがあれば、このような半角候補文字を含む英単語候補は、英単語候補から除外される。

【0043】上記のテーブルに記憶された標準条件は、たとえば、下記表1に示すようなものである。すなわち、変化点の数num＝1の場合には、当該半角候補文字の位置条件として、分離度SPが1の場合に、半角英数字であるものと判定される。また、変化点の数num＝2の場合には、位置条件として分離度SP＝1である場合と、位置条件として分離度SP＝2を満たし、かつ、画素条件としてB（2）＞B（1）×4を満たす場合とに、当該半角候補文字が英数字であるものと判定される。さらに、変化点の数num＝3の場合において、当該半角候補文字が英数字であると判定されるのは、位置条件として分離度SP＝1かつSU+W（4）＞Vi×0．09を満たし、さらに画素条件として（Vi－W（1）－W（4））＞Vi×0．7を満たす場合である。さらに、変化点の数num＝4の場合には、分離度SP＝1であって、かつSU＞V×0．26なる位置条件が満たされる場合に、半角候補文字が英数字であるものと判定される。なお、図14に示されているように、Viは囲み切出矩形35の縦幅であり、Vは行幅である。

【0044】

【表1】

13　　14

num	画　素　条　件	位　置　条　件
1	—	SP=1
2	—	SP=1
	B(2)>>B(1)*4	SP=2
3	(Vi-W(1)-W(4))>Vi*0.7	SU+W(4)>Vi*0.09 SP=1
4	—	SU>V*0.26 SP=1

【００４５】この照合の結果、英単語候補を構成する半角候補文字の全てが、上記表１の標準条件を満たしていれば、このような半角候補文字列は英単語候補として確定される。その一方で、英単語候補を構成するいずれか１つの半角候補文字が上記表１の標準条件を満たさない場合には、このような文字列は英単語候補でないものとして、英単語候補から除外される。

【００４６】図１５は、英単語候補を抽出するための処理を説明するためのフローチャートであり、図４のステップｎ７～ｎ９の具体的な処理が示されている。ステップｒ１では、パラメータｉに１が代入され、別のパラメータＣＯＵＮＴに０が代入される。次にステップｒ２では、パラメータｉが当該行内の文字要素の数に達したかどうかが判断され、文字要素の数に達していれば処理を終了する。パラメータｉが文字要素の数に達する以前であれば、ステップｒ３において、当該行を構成するｉ番目の文字要素が半角候補文字であるかどうかが判断される。半角候補文字でないときには、ステップｒ１３でパラメータｉがインクリメントされた後にステップｒ２に戻り、半角候補文字であるときにはステップｒ４進む。

【００４７】ステップｒ４では、ｉ番目の文字要素と（ｉ＋１）番目の文字要素との間隔が変数Ｓに代入され、また、ｉ番目の文字要素と（ｉ＋１）番目の文字要素とのピッチが変数Ｐに代入される。さらに、パラメータＣＯＵＮＴに１が代入される。そしてステップｒ５でパラメータｉがインクリメントされた後に、ステップｒ６では、インクリメントされた後のｉに従って、ｉ番目の文字要素は半角候補文字であるかどうかが判断される。半角候補文字でなければステップｒ１３に進み、半角候補文字なら、ステップｒ７で上記の画素特徴などが抽出される。

【００４８】次にステップｒ８では、抽出された画素特徴などが、上記表１に示されたテーブルと照合される。ステップｒ９では、①変数Ｓと、ｉ番目の文字要素と（ｉ＋１）番目の文字要素との間の間隔とがほぼ等しいかどうか、②変数Ｐと、ｉ番目の文字要素と（ｉ＋１）番目の文字要素とのピッチがほぼ等しいかどうか、③ｉ番目の文字要素は行末の文字要素かどうかが、それぞれ判断される。そして、上記①～③のいずれか１つについて肯定的な判断がなされると、ステップｒ１０でパラメータＣＯＵＮＴがインクリメントされてステップｒ５に戻る。また、上記①～③のすべてについて否定的な判断がなされたときには、ステップｒ１１に進む。

【００４９】ステップｒ１１では、パラメータＣＯＵＮＴが３よりも大きいかどうかが判断れ、３よりも大きければ（すなわちＣＯＵＮＴ＝４となると）、ステップｒ１２において、（ｉ－ＣＯＵＮＴ）番目からｉ番目の文字要素が英単語を構成する文字要素であるものとされる。この後の処理はステップｒ１３に進む。なお、ステップｒ１１で、パラメータＣＯＵＮＴが３以下である場合には、ステップｒ１２を経ずにステップｒ１３に進み、次の文字要素についての処理が行われる。

【００５０】このようにして、一定間隔または一定ピッチで一定数以上の半角候補文字が連続しており、かつ各半角候補文字が半角英文字であると判断されるときに、このような半角候補文字列が英単語候補として抽出されることになる。再び図４を参照して、英単語候補が抽出された後の処理を説明する。ステップｎ１０では、切出位置の修正が行われる。すなわち、英単語候補を構成する半角候補文字については、ステップｎ３での文字要素の切出位置をそのまま用いればよいが、英単語候補を構成しない半角候補文字は、全角分離文字の一部をなす文字要素である可能性が高いため、切り出すべき画像の再構成が必要となる。すなわち、隣接する半角候補文字同士を結合させて切り出すために、切出位置の修正が行われるのである。

【００５１】次に、ステップｎ１１では、文字切出部６０（図１参照）において１文字ごとの画像である文字画像が切り出され、この文字画像と英数字辞書６３または認識用辞書６４（図３参照）を参照することで、文字認

15

識処理が行われる。すなわち、切り出された文字画像から特徴量が抽出され、この抽出された特徴量と辞書６３，６４に記憶された各文字の基準となる特徴量との類似度が計算される。そして、類似度が最大となる文字の文字コードが、認識候補としてその類似度とともに出力部６５から出力される。

【００５２】このような認識処理において、類似度算出部６２では、英単語候補を構成する半角候補文字については、英数字辞書６３を参照し、その他の文字要素については認識用辞書６４を参照する。したがって、英単語候補を構成する文字要素については、類似度計算の対称となる文字が半角英数字に限定されるので、類似度計算が格段に簡単に行える。これにより、認識処理が高速化されることになる。なお、英単語候補を構成する半角候補文字については、英単語を構成する文字を認識するための既存のあらゆる手法を適用できる。

【００５３】ステップｎ１２では、認識候補の類似度が所定値を超えているかどうかが判断され、認識候補の類似度が著しく低いとき、たとえば平均類似度の１割未満であるときには、切出ミスであると判断され、ステップｎ１２からステップｎ１０に戻って切出位置の修正が行われる。すなわち、切出ミスとされた文字要素が英単語候補を構成する半角候補文字であるときには、この文字要素が全角分離文字の一部を構成する文字要素であるものとして切出しがやり直される。また、全角分離文字の一部を構成するものとされた文字要素に関して切出ミスと判断されたときには、当該文字要素が半角文字であるものとして、文字画像の切出が再度行われる。このようにして、再度切り出された文字要素に関して、上記の文字認識処理が行われることになる（ステップｎ１１）。

【００５４】以上のように本実施例の光学的文字読取装置では、認識処理を行うことなく半角英文字からなる英単語候補を抽出することができるから、文字の切出を高速に行うことができる。しかも、英単語候補の抽出には、日本語文書中に含まれる英単語のレイアウト上の特徴のほか、各半角候補文字の位置特徴や画素特徴を参照して行われているから、文書フォーマットの異なる種々の文書に良好に対応して、英単語候補を正確に抽出することができる。したがって、文字の切出しを文書フォーマットによらずに正確に行わせることができ、この結果、文字認識処理の高速化に寄与することができる。

【００５５】また、文字認識の過程で英単語候補が抽出されるので、たとえば１文字ずつの文字認識の後に、英単語のスペリングを調べるスペルチェックなどの処理を容易に行うことができるという利点がある。なお、本発明は上記の実施例に限定されるものではない。たとえば上記の実施例では、英単語候補の抽出の際に、所定数以上の半角候補文字列の後ろに所定長以上の余白が形成されていることを条件としたが、このような条件の代わりに、所定数以上の半角候補文字列の前に所定長以上の余

16

白が形成されていることを条件としたり、所定数以上の半角候補文字列の前と後ろの両方に所定長以上の余白が形成されていることを条件としたりしてもよい。

【００５６】また、上記の実施例では、画素特徴の１つとして、半角候補文字の行方向に関する中心位置での縦方向への走査において、白画素→黒画素→白画素のように変化する変化点の数ｎｕｍを採用しているが、このような画素特徴の代わりに黒画素→白画素→黒画素のように変化する変化点の数を採用してもよい。さらに、上記の実施例では、白地に黒色の文書画像が形成されている場合を例に採ったが、黒地に白色の文書画像が形成されている場合についても、本発明は容易に応用することができる。すなわち、この場合には、白画素の塊を文字要素として抽出することになる。

【００５７】さらに、上記の実施例では、文書画像は、原稿を光学的に読みとるようにして画像メモリ４に入力されているが、文書画像を表すデータを通信回線を介して取得し、この取得した画像データを画像メモリに記憶させる構成としてもよい。その他、本発明の要旨を変更しない範囲で種々の設計変更を施すことが可能である。

【００５８】

【発明の効果】以上のように本発明の文字認識装置によれば、日本語文書中に混在する半角英文字は英単語を構成していることが多いことを利用して、英単語候補を抽出することによって、文字画像の切出しの最適化が図られている。これにより、半角文字は当初から半角文字として切り出し、全角分離文字は当初から半角候補文字同士を再結合させた状態で切り出すことができるから、文字切出処理を正確に行うことができる。しかも、英単語候補の抽出は、文字認識処理を要することなく行われるから、正確な切出処理を高速に行うことができ、したがって文字認識処理を高速化することができる。

【００５９】さらに、英単語候補の抽出に当たっては、文書画像のレイアウトの特徴のほか、英単語候補を構成する各半角候補文字の位置特徴や画素特徴をも参照しているので、文書フォーマットの相違に依らずに英単語候補の抽出を良好に行うことができる。これにより、切出処理の一層の最適化が図られるから、文字切出処理を極めて正確に行うことができ、文字認識処理の回数を低減して、全体の認識処理に要する時間を短縮することができるようになる。

【図面の簡単な説明】

【図１】本発明の文字認識装置の一実施例である光学的文字読取装置の要部の構成を示すブロック図である。

【図２】光学的文字読取装置の全体の構成を示すブロック図である。

【図３】認識部の詳しい構成を示すブロック図である。

【図４】本発明の一実施例の文字認識装置による文字認識処理を説明するためのフローチャートである。

【図５】１行毎の画像の切出処理を説明するための図で

ある。

【図６】文字要素毎の画像の切出処理を説明するための図である。

【図７】各文字要素の周辺分布を求めた例を示す図である。

【図８】半角候補文字の抽出処理を説明するための図である。

【図９】半角候補文字の抽出処理を説明するためのフローチャートである。

【図１０】日本語文書中に混在している半角英文字の印字特徴を説明するための図である。

【図１１】英単語候補の抽出処理を説明するための図である。

【図１２】画素特徴の抽出処理を説明するための図である。

【図１３】画素特徴の抽出処理を説明するための図である。

【図１４】画素特徴を説明するための図である。

【図１５】英単語候補の抽出処理を説明するためのフローチャートである。

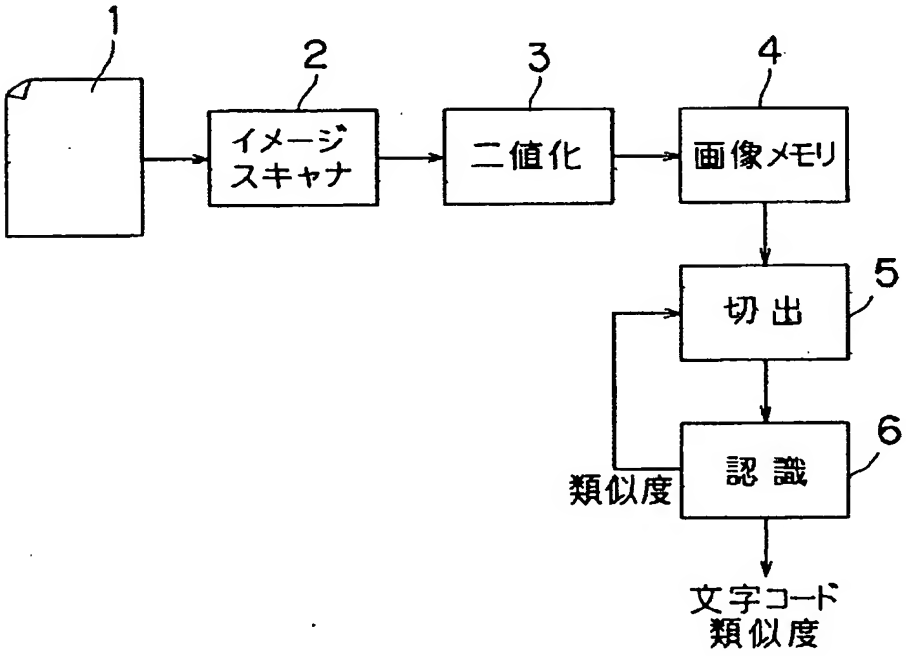
【図１６】従来技術の問題点を説明するための図である。

【図１７】他の従来技術の問題点を説明するための図である。

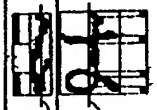
【符号の説明】

５　切出部
６　認識部
５１　行切出部
５２　仮切出部
５３　文字サイズ推定部
５４　半角候補文字抽出部
５５　英単語候補抽出部
５６　画素特徴抽出部
５７　照合部
５８　画素特徴記憶部
５９　英単語候補修正部
６０　文字切出部
６１　特徴抽出部
６２　類似度算出部
６３　英数字辞書
６４　認識用辞書
６５　出力部

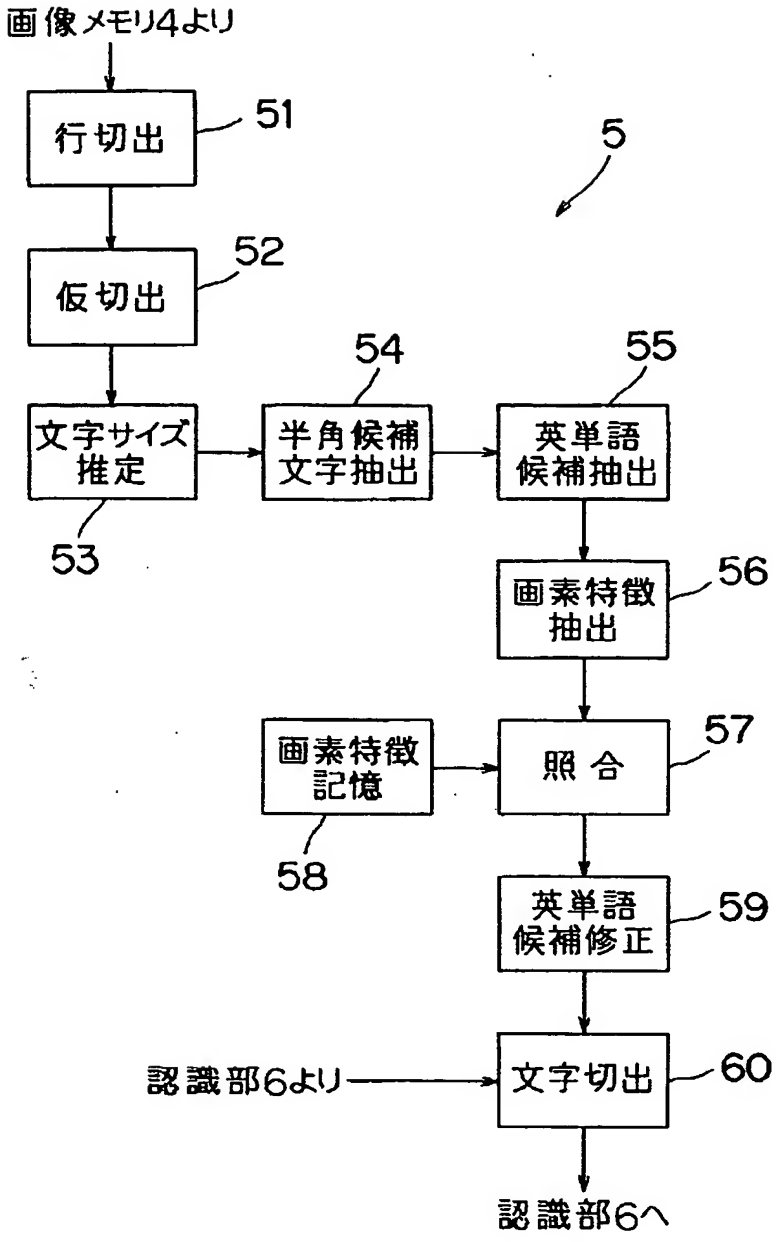
【図２】



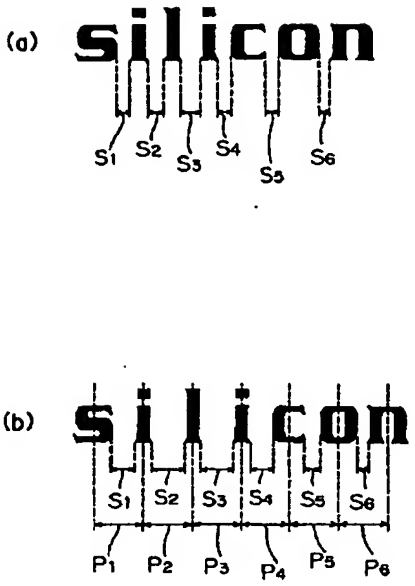
【図１３】

	(a)	(b)
		
num	2	4
W(1)	0	0
W(2)	15	10
W(3)	0	14
W(4)	0	5
W(5)	0	0
B(6)	16	2
B(7)	12	4
B(8)	0	2
B(9)	0	4
B(10)	0	0
SP	1	1

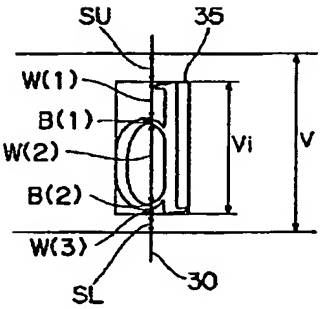
【図1】



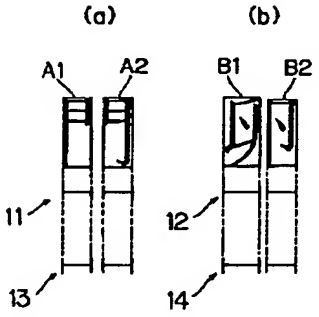
【図10】



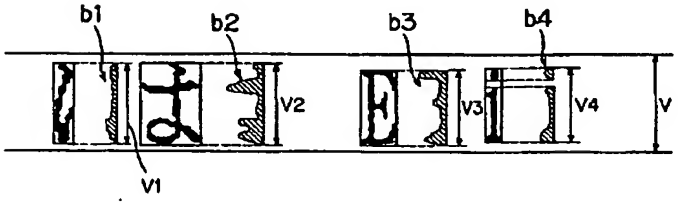
【図14】



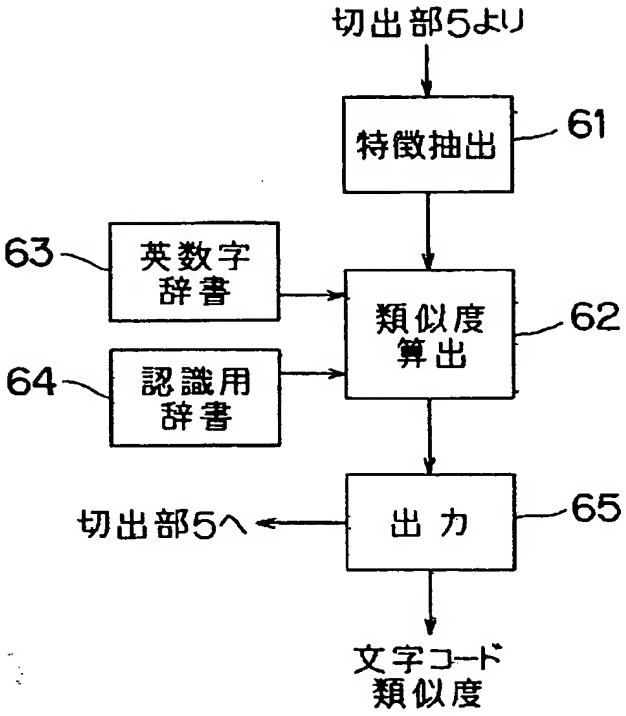
【図16】



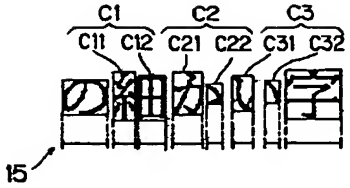
【図7】



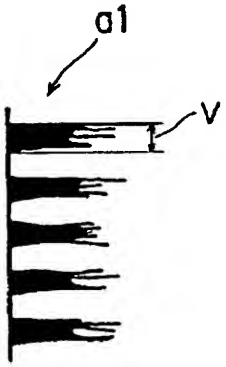
【図３】



【図１７】

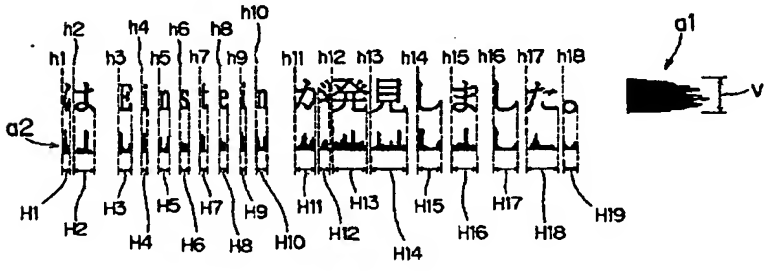


【図５】

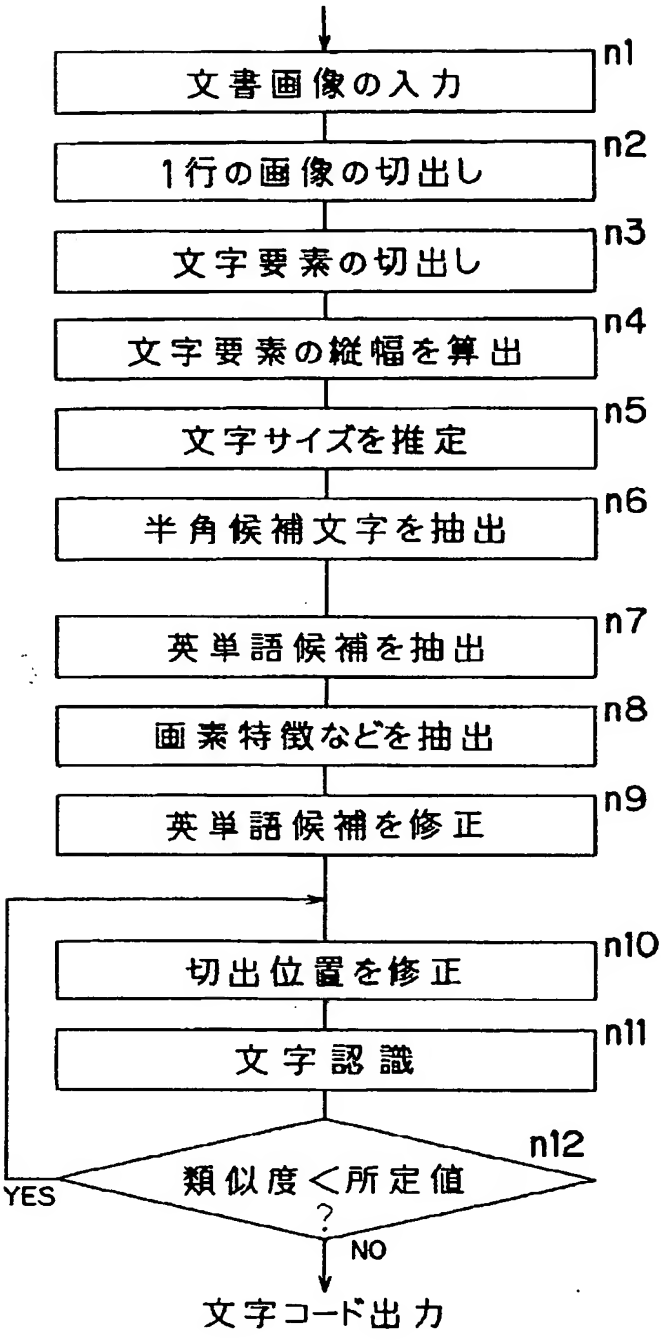


L1 →ンサ系が採用される．ここでは 10 個の変位情報すべてを生か
L2 →して制御に必要かつ十分な５個のデータに処理するが，これは
L3 →電磁石吸引力に正確に対応するデータでなくてはならず．この
L4 →場合は吸引力作用点の変位を算出し，これをフィードバックす
L5 →る．これらと入力指令値との差は PID 演算，電力増幅器を経て

【図６】



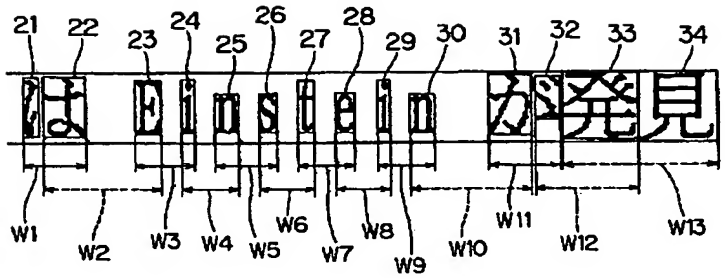
【図４】



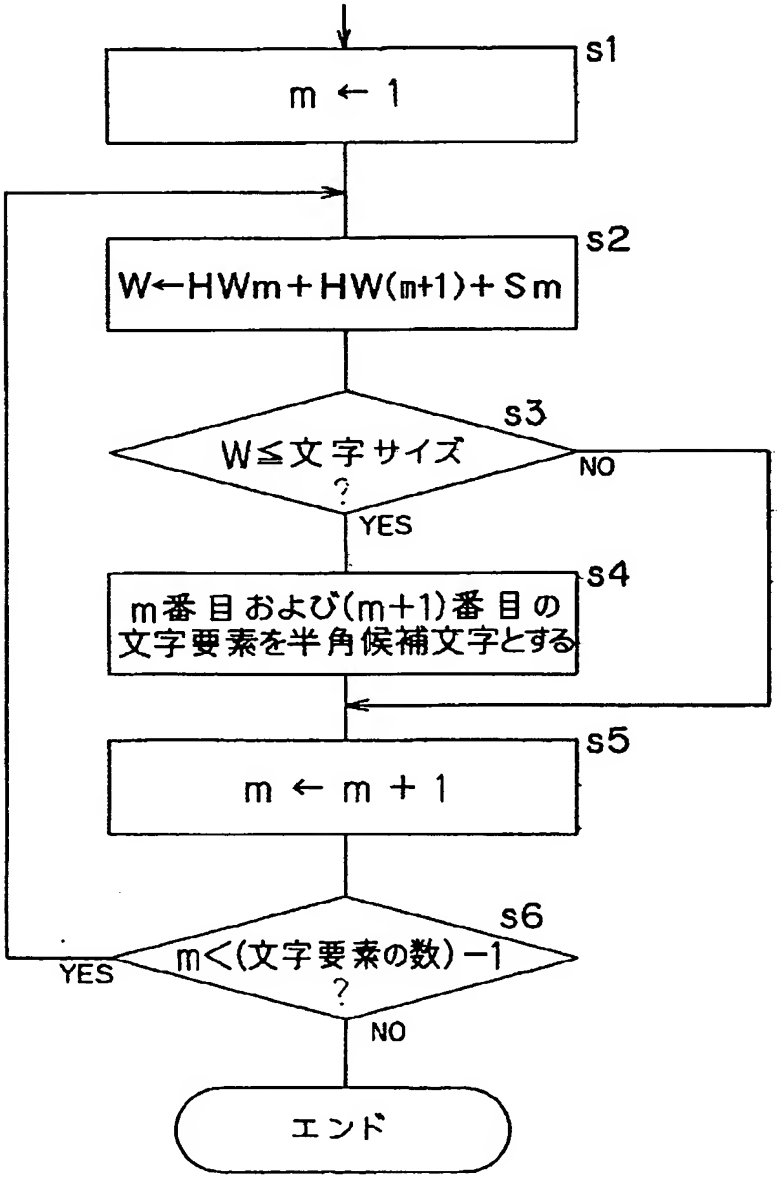
【図１２】

	(a)	(b)	(c)	(d)
	C1 d1 C2 C3 30	d2 30	d3 30	d4 30
num	3	2	1	3
W(1)	0	0	0	0
W(2)	14	3	24	8
W(3)	15	0	0	8
W(4)	0	0	0	0
W(5)	0	0	0	0
B(6)	2	7	4	3
B(7)	4	28	0	4
B(8)	2	0	0	3
B(9)	0	0	0	0
B(10)	0	0	0	0
SP	1	2	1	1

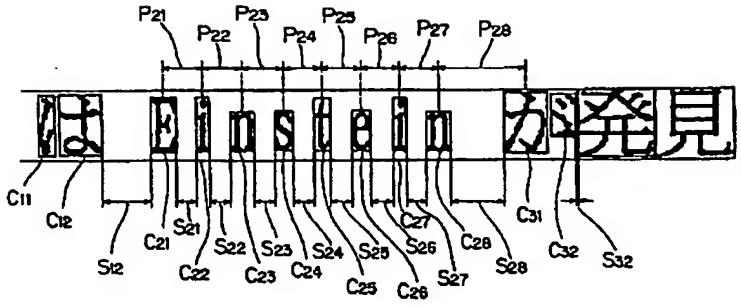
【図８】


21
22
23
24
25
26
27
28
29
30
31
32
33
34
W1
W2
W3
W4
W5
W6
W7
W8
W9
W10
W11
W12
W13

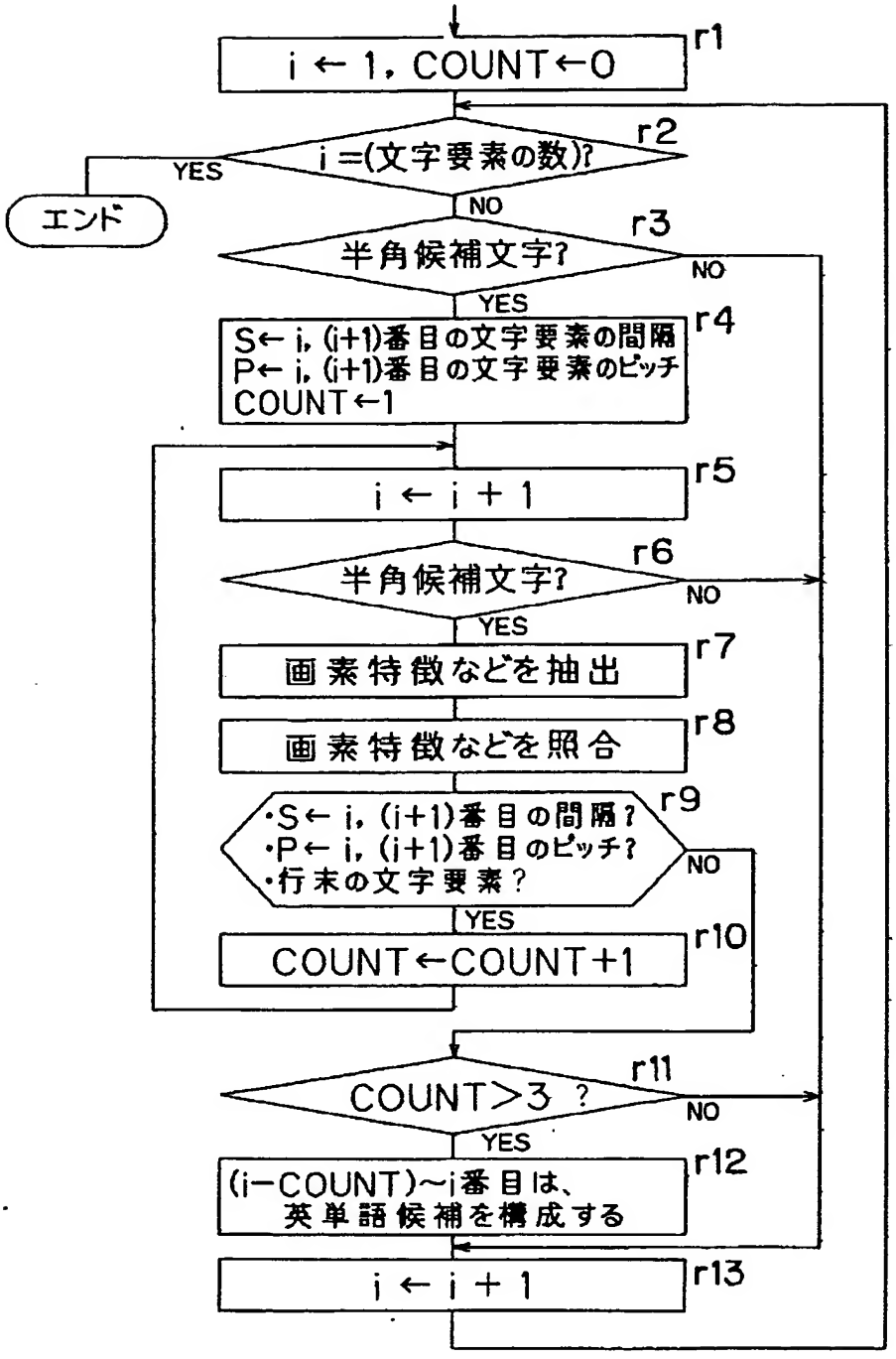
【図９】


s1
m ← 1
s2
W←HWm＋HW(m+1)＋Sm
s3
W≦文字サイズ
?
NO
YES
s4
m番目および(m+1)番目の
文字要素を半角候補文字とする
s5
m ← m ＋ 1
s6
m<(文字要素の数)－1
?
YES
NO
エンド

【図１１】


P21
P22
P23
P24
P25
P26
P27
P28
Einstein
C11
C12
S12
C21
S21
C22
S22
C23
S23
C24
S24
C25
S25
C26
S26
C27
S27
C28
S28
C31
C32
S32

【図15】


i ← 1, COUNT←0
r1
i =(文字要素の数)?
r2
YES
エンド
NO
半角候補文字?
r3
NO
YES
S← i, (i+1)番目の文字要素の間隔
P← i, (i+1)番目の文字要素のピッチ
COUNT←1
r4
i ← i + 1
r5
半角候補文字?
r6
NO
YES
画素特徴などを抽出
r7
画素特徴などを照合
r8
・S← i, (i+1)番目の間隔?
・P← i, (i+1)番目のピッチ?
・行末の文字要素?
r9
NO
YES
COUNT←COUNT+1
r10
COUNT>3 ?
r11
NO
YES
(i−COUNT)～i番目は、
英単語候補を構成する
r12
i ← i + 1
r13